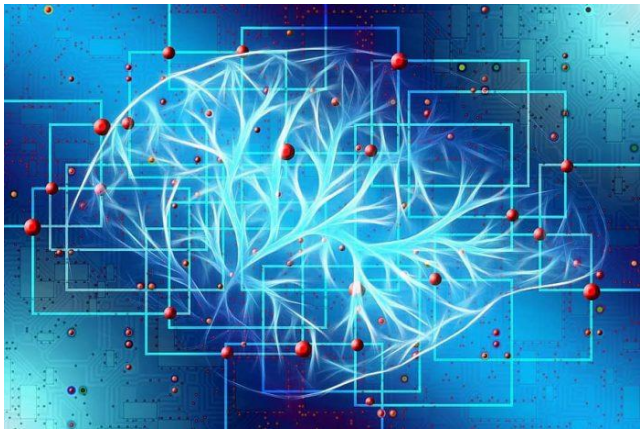
Blog of the APA

A Neurophilosophy of Governance of Artificial Intelligence and Brain-Computer Interface

By Nayef Al-Rodhan

June 1, 2020



This is post four in a short-term series by Prof Nayef Al-Rodhan titled "Neurophilosophy of Governance, Power and Transformative Innovations." This series provides neurophilosophical perspectives and multi-disciplinary analyses on topics related to power and political institutions, as well as on a series of contemporary transformative technologies and their disruptive nature. The goal is to inspire innovative intellectual reflections and to advance novel policy considerations.

When Kant wrote that moral judgment derives from rationality – and the autonomy of rational will – he did not offer a perfectly clear idea of what reason meant. However, in his conceptualization, neither Kant, nor other philosophers in the rationalism vs empiricism debate, could foresee the interference of machines with the mind, and the implications of connecting or synching technologies to the human brain.

While many factor interfere with 'rationality' (a term that neurophilosophy has re-evaluated in new lights in recent years) the advent of artificial intelligence, neural interfaces, and especially technologies such as 'closed-loop' systems (which record brain signals and deliver stimulation in response), raise unique questions for philosophy.

Neural interfaces connect the brain or parts of the nervous system to digital or IT systems and devices without manual input (e.g. keyboard, joystick etc.). There are currently two main types of neural interfaces: those that 'read' the brain, thus recording brain signals and decoding their meaning, and those that 'write' to the brain by stimulating or manipulating activity in specific regions.

The original intended use of brain-computer interfaces was purely medical and was proposed to patients suffering from neuronal dysfunction like stroke, epilepsy, paralysis, parkinsonism, depression etc. The most well-known example of implantable interface today is the cochlear implant –estimated to be worn by about 400,000 people worldwide, and which enables them to hear despite damage to the cochlea. Other, more sophisticated systems, such as Neuralink's ultra-high bandwidth brain-computer interface (BCI) could record signals from the brain to an unprecedented extent compared

to anything currently on the market – even to the point of providing a sense of touch to neuroprosthetic movement control.

Increasingly, however, the uses of such interfaces – sometimes called electroceuticals – is expanding well beyond medicine, to gamers who use headsets to control on-screen characters, to students to act as concentration aids, and even to some companies reportedly using technology to monitor the mood of their employees.

As uses expand, so do risks: the risks of loss of privacy and autonomy, the risks of thoughts and intent being accessed by companies, or the risks of widening inequalities in society. I wrote about some of the inherent risks related to enhancement technologies in a previous post and BCIs subscribe to many of the same concerns. In this post, I wish to refer to a range of specific ethical and philosophical aspects of BCI and further on, following the merging of BCI with artificial intelligence.

From medical use to enhancement

The purely medical uses of neural interfaces, are not only easy to defend, they could be even seen as a moral imperative: technology and devices that help restore some lost functions and in the process, a life of less pain and more autonomy to individuals, should be pursued assiduously.

Even more so, as many medical conditions are currently still partially drug-resistant (for example, about 20-30% of epilepsy is considered to be drug-resistant), electroceuticals could be a much-needed and more effective way of helping patients. That is not to say that the medical applications come free of risks or ethical concerns. For example, one unwanted consequence of BCIs can be psychological harm, following unmet expectations in some patients and care-givers whose hopes to recover some lost functions with a BCI fail to realize. Outside the scope of medical uses, brain-computer interfaces raise different questions.

Philosophically, uses of BCI that lead to cognitive enhancement, or to the potential of operating computer-controlled devices raise challenges to the notion of agency, personal autonomy, as well as the core of the meaning of trust.

One critique to neural interfaces has pointed to the risk of "neuro-essentialism", which is the idea that the brain is the defining and essential part of a person and that everything about the human experience can be explained by and attributed to a range of neurochemical and neuroanatomical reactions. In other words, subjectivity, the self, and human identity become reduced to the *brain* and what techniques such as functional magnetic resonance imaging (fMRI) reveal about it. Or, this oversimplification overlooks the fact that neuroscience does not take the mind to be 'static', but it accounts for a wide array of factors that impact our cognition and decisions. We are born as *predisposed tabula rasa*, with no predefined notions of good or bad; we are only minimally equipped with a predilection for survival (and seeking actions that maximize our chances of survival). Everything else, including our moral compass, is impacted by the environment and conditions in our surroundings.

However, while the critique of neuro-essentialism may be misplaced, there remain very valid ethical and philosophical concerns about BCIs.

In the context of increased miniaturization of interfaces, these quagmires only multiply. For example, "neural dust", a tiny, miniaturized form of brain-machine interface, is smaller than a grain of rice and could be inserted in the body to stimulate nerves, muscles or organs in real-time. It would be powered by ultrasounds, and thus penetrate almost anywhere in the body (unlike radio waves). Apart from some medical risks, these sensors do not really pose outstanding ethical challenges when used, for

instance, for the treatment of epilepsy or the treatment of type 2 diabetes. However, the long-term prospect indicates a much broader range of uses, including eventually implanting them in the central nervous system (not only the peripheral nervous system and muscles). When such sensor motes would be used, for instance, to stimulate the immune system, or to stimulate cognitive functions, the ethical challenges become more complex.

Given their implantable nature and extremely small size, and therefore, invisibility to one's peers, such devices may be worn for a long time without being aware that one's colleagues, friends or superiors are effectively enhanced and at a comparative advantage, physically and/or cognitively. If and when others learned about the use of neural interfaces, a key component of human social life will be severely affected, namely trust. Trust is not only foundational to human existence, it pervades all aspects of life and relationships, including social, political and economic institutions. In neuroeconomics, trust has been valued as a fundamental element in management, with high returns on performance and productivity (a Harvard study revealed a 50% increase in productivity, 74% less stress, 106% more energy at work).

Every aspect of human life is impacted by trust or conversely, by lack or distortions of trust. Neurochemically, trust is shaped by oxytocin, a neuropeptide produced in the hypothalamus and stored in the posterior pituitary. Oxytocin plays important roles in social behavior, including in bonding, forming social attachments, and in the regulation of neuroendocrine responses to social stressors and anxiety: trust is important in social interactions, allowing for the 'risky first step' towards another individual.

Some subcortical brain structures such as the amygdala and brainstem effector sites are also known to be involved in trusting behaviors; these structures normally process fear, danger as well as the risk of social betrayal. Other studies showed that in situations of threat or betrayal, neuronal connections are quite literally suppressed or compromised as a result; for instance, a study showed that aversive affect and the existence of a threat (not being able to trust others is conducive to a sense of threat) prevent connections between the amygdala and the temporoparietal junction (which plays a role in mentalizing), and thus the capacity to mentalize about others or to evaluate others' emotions are diminished. The existence, absence or distortion of trust thus play fundamental roles in decision-making, how we relate to others but also to our own well-being: being able to trust others is highly gratifying, as well as engaging a host of circuitries in the brain that encourage further social bonding. It is thus not far-fetched to posit that trust maximizes – and is perceived as maximizing – our chances of survival.

The neuroscience of (the ability to) trust and of betrayal point not only to their foundational roles in establishing or severing social ties but also to possible consequences of uses of BCIs, in particular when BCIs are used as means of enhancement. Deception, a sense of betrayal or unfairness would accompany the realization that someone is relying on a neural interface to stay fit, focused or alert for longer. Making 'moral choices' (which Kant had attributed to rationality) when lacking trust in others or under the suspicion of deceit becomes more difficult.

From interface to a 'part of you'

In a future when neural interfaces develop a closer and more symbiotic relationship with artificial intelligence, their potential would become further heightened and the implications even more unsettling. Several interfaces today already rely on AI to read and convert neural signals into digital data, or to decode some of the neural commands sent by the brain. What if such brain signals were tracked not only to move a prosthetic arm, but also to detect fatigue levels and prompt an

intervention? The sense of connection to the device may become much deeper, to the point that the interface became indistinguishable from the individual. (Lawmakers will probably grapple with this as well. Currently, prosthetics are treated under property law, yet an interface so deeply connected to the individual to the point that it becomes part of their sense of self may need to fall under some other legal regime.)

Interfaces that will link human thought with AI and machine-learning, allowing better predictions and better courses of future actions, could, in theory, allow for enhanced decision-making capabilities, improved situational awareness and perhaps even new sensory experiences. While still mostly premised on medical uses, the market for linking neural interfaces with AI is expanding every year and so does the non-medical range of uses. For example, a wearable interface called AlterEgo, premised on the idea of "silent speech", would allow humans to communicate silently and in a concealed manner with machines, artificial intelligence systems and perhaps even other people, quite literally without having to open their mouth or performing any externally observable movements. This natural user interface (NUI) would work by receiving feedback through audio and via bone conduction, just as one would be talking to oneself. Currently purported for medical use, e.g. for patients with speech impairments, its uses and appeal may transcend the medical sphere to become tools for extremely fast access to information and for effectively weaving AI and computers into the human personality, like a "second self" – in fact, *inter alia*, the developers of AlterEgo hope to achieve exactly that.

A first identifiable risk concerns the possibility of hacking of such devices. 'Brainjacking', meaning the possibility that attackers may exert malicious control over brain implants, is deemed likely both in the form of 'blind attacks', requiring no patient-specific knowledge (e.g. cessation of stimulation, inducing damage in the tissue or theft of information), or targeted attacks in the form of impairment of motor function, modification of emotions or affect, induction of pain, or even modulation of the reward system.

Even without such risks materializing, and assuming such interfaces were hack-proof, the merging of interfaces with AI systems is clearly going to redefine our perception of 'normality', as well as of personal autonomy and identity. That is because, even in more rudimentary forms, BCIs impact the brain, a powerful reminder that the use of such technologies needs to be strictly regulated.

More recent evidence has shown that brain plasticity occurs after just one hour of BCI – for example, the same study showed increased intensity of grey matter in occipital/parietal areas, among other changes. There is a wealth of neuroscientific evidence pointing to the fact that the adult human brain is capable of structural reorganization, and that engagement in certain long-term activities leads to changes in grey-matter-density or volume in certain regions of the brain. However, in the case of BCIs, brain plasticity occurs in a matter of a few hours or weeks – though it should be noted that the locations of these changes do not seem to be necessarily occurring in the same places as for long-term changes. The question of authenticity of human existence will, however, be inevitable going forward, as will the notion of the authenticity of free will.

Furthermore, it is important to realize these impacts are not transient, and they may contribute to deeper epigenetic changes. Studies in epigenetics, which is a discipline that looks into heritable changes in gene expression (not involving alterations in the underlying DNA sequence), typically consider "lifestyle", broadly speaking, and individual genetic background as intertwined. Environmental factors may influence epigenetic mechanisms such as DNA methylation, histone modifications and microRNA expression. These flexible genomic parameters can change under many types of exogenous influences or as a result of many diseases – traumatic experiences too have been linked to certain genome alterations. Importantly, in the long run, these may be passed on to one's

offspring. The long-term and profound consequences of BCI should thus also be judged from this perspective. Many factors in our environment already interfere with or alter our genes, but, at the very least, we have been warned repeatedly about the negative consequences of many of these factors, such as pollutants, tobacco or alcohol consumption. The epigenetic consequences of BCI are yet to be studied but, regardless, users and consumers need to be informed that long-term impacts may occur.

Benefits and risks: the importance of regulation

We can surely accept that an individual (and their brain) changes as a result of experience and engagement with the world, but a more immediate, or sudden change occurring hours after the interference of a device is a more difficult reality to accept. In many ways, this is reminiscent of Nozick's experience machine, which I quoted previously in another discussion of enhancement and artificial neuromodulation. Such a change is not only hard to accept on grounds that it is not 'natural', but also because it makes it harder to understand to what extent there is something underlyingly authentic left about the individual or if they are now (following the interaction with BCI) a 'new' person.

Though at varying degrees, neuro-ethicists present in clinical settings have noticed many instances of BCIs leading to personality changes, following applications of deep brain stimulation for the treatment of Parkinson's diseases (first approved by the FDA in the US in 1997). Moreover, in many instances, the changes may affect a person's own perception of themselves, and another question is whether the person who had undergone a stimulation with an interface could reflect on how they had changed. This does not mean we should regard all BCIs are menacing. As Liam Drew puts it: "to observe a person with tetraplegia bringing a drink to their mouth using a BCI-controlled robotic arm is spectacular."

The real challenges come when a closed-loop system is merged with machine learning software, which learns to analyze data and generate algorithms, and effectively 'takes over' part of the decision-making of an individual. It is effectively like inserting a decision-making device into someone's brain. Again, the difference between some therapeutic uses vs enhancement/neuromodulation exposes the real challenge of loss of agency: a device that monitors blood glucose and learns to automatically control insulin release is a form of decision-making on behalf of someone that is hardly controversial, in fact welcome. But, if a similar device is inserted for mood disorders, it may prevent someone from experiencing negative emotions even when such emotions are normal, such as at a funeral.

This leads to a final point, which concerns the critical urgency of regulating BCIs, and especially those marketed for non-medical purposes. Human nature is fragile and malleable, and regulation is in our best 'existential interest'. I previously theorized that human nature is *emotional amoral and egoistic.* (See a previous post for a more elaborate discussion) We are deeply *emotional* beings, and as a wealth of evidence shows, emotionality, cognition and learning, and decision-making are tightly connected in the brain. We are also *amoral*, in the sense that we are born neither innately moral, nor immoral, but rather amoral and our moral compass will be greatly influenced by personal and political conditions in our environment. The only minimal genetic predispositions we have are towards survival of the self (and of our kin), which is a basic form of *egoism*. Given these fundamental elements that define our nature, regulation of BCIs is critical. Left to our own 'best judgement', the uses and abuses of such technologies will only become rampant, with disastrous consequences in the long run.

Apart from the frailty and malleability of our nature, regulation is also important as a means of preventing excessive accruement of power and data by private actors. Quite obviously, the minimal oversight that now defines consumer technology can only lead to disempowerment of

consumers. This is will be critical going forward, for our individual and collective dignity, and also for our understanding of free will, our freedoms, authenticity and autonomy, in a highly interactive and intrusive future digital world.

Nayef Al-Rodhan

Prof. Nayef Al-Rodhan (@SustainHistory) is a Neuroscientist, Philosopher and Geostrategist. He is an Honorary Fellow at St Antony's College, University of Oxford, and Senior Fellow and Head of the Geopolitics and Global Futures Programme at the Geneva Centre for Security Policy, Geneva, Switzerland. Through many innovative books and articles, he has made significant conceptual contributions to the application of the field of neurophilosophy to human nature, history, contemporary geopolitics, international relations, cultural studies, future studies, and war and peace.